MECH60017/MECH96014/MECH96038 Statistics

Coursework

INSTRUCTIONS

Hand-out Date: 28 February 2022 11am **Hand-in Date**: 14 March 2022 11am

The final report must be typed up and should be a properly structured document in PDF. Answers in your report should be written in complete sentences. Complete every question, address each point clearly by providing the mathematical expressions if required, and MATLAB code for each computation, along with your comments. Marks will be given for quality of presentation. Your report should not exceed five A4 pages. You may include a cover page and appendices and these are not included in the five-page limit; references (if any) should be included in the five pages.

Submit (a) a PDF version of your report, and (b) your MATLAB code and output to Blackboard by the deadline stated above (there are two submission boxes in the Blackboard folder *Coursework Information and Submissions*). I should be able to execute your code, without making any modifications, except a first line to read in the data. Badly formatted or unclear code and output will be penalised. Once the report is uploaded there is option for re-uploading, but if possible, avoid last minute uploads as the system can crash if it simultaneously receives too many requests.

This coursework will be scored out of 20. A total of 4 marks are available for presentation. This is worth 10% of the final grade.

IMPORTANT: at the very beginning of your MATLAB code, add and run the command `rng(CID)`, where `CID` is your College ID number. This sets the seed for the random number generator in MATLAB so that the output produced by your MATLAB code can be reproduced exactly.

Modelling Fuel Efficiency

The purpose of this coursework is to develop your practical skills in statistical modelling, using data extracted from a real application.

Your task is to model the fuel efficiency of vehicles using linear regression and clustering.

You are provided with your own, unique datasets. For questions 1-4, use your training dataset and for question 5, use your test dataset. To access your datasets go to the following web address:
`https://imperiallondon-my.sharepoint.com/personal/ipapatso_ic_ac_uk/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fipapatso%5Fic%5Fac%5Fuk%2FDocuments%2FData`
The datasets have been named after your CID number, i.e. if your CID number is 1327290, download the files `train1327290` and `test1327290` and save the files as csv files in a suitable location. You can then import the files as usual into MATLAB.

If you have difficulty downloading your dataset, please email me. For any other issues, please use Blackboard's discussion board.

You should see six columns:
`l100`: fuel efficiency, measured in litres per 100km
`t100`: time to go from 0 to 100km/h, measured in seconds
`mass`: mass of the vehicle, measured in kilograms
`fuel`: fuel type (petrol, diesel)
`colour`: colour (white, other)
`displacement`: engine size, measured in litres

Questions

1. **Exploratory data analysis**

   (a) Compute the mean, median, standard deviation, and interquartile range of the fuel efficiency and present these in a table. Plot a histogram of the fuel efficiency. Which of the above measurement(s) would you use to best describe the data and why?

   (b) Construct boxplots of the fuel efficiency by categorical variables and a matrix of clearly-labelled scatterplots of all pairs of continuous variables. Comment on these

plots.

[You should use the exploratory information to help you to decide in what form the variables are included in the model (question 2) to satisfy the assumptions of linearity/additivity, e.g. are any transformations necessary? Before proceeding, recode the categorical variables as numerical (binary 0/1), e.g. recode `fuel` as `petrol` equal to 0 for diesel and 1 for petrol vehicles; recode colour too. Use the binary variables in the modelling section.]

2. **Modelling**

   (a) Fit a simple linear regression model for the fuel efficiency versus the mass, using the `fitlm` command in MATLAB. Make a clearly-labelled scatterplot and plot the linear fit on your scatterplot. Comment on the appropriateness of this model for your data.

   (b) Some ways to assess the fit of multivariable linear regression models $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where the $\epsilon_i \sim N(0, \sigma^2)$, are the multiple coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

   the mean squared error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

   and Akaike's Information Criterion (AIC):

$$AIC = 2 * q - 2 * l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$$

   where $l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ is the log-likelihood of the model at the maximum and q is the number of parameters estimated in the model.

   Research and briefly explain what the above measures tell you and how they can be used to select between models.

   (c) Find the model that you think best predicts fuel efficiency. You may consider including any of the predictors or parametric functional forms of these. You may also consider including pairwise interaction terms. A criterion for including interaction

terms is to try including an interaction if the main effects are large, and keep it in if this improves the model fit. Summarise your findings in a table with one row for each model considered, including R^2, MSE and AIC for each model.

3. **Model checking and interpretation**

   (a) For your chosen model from question 2 (c), calculate the residuals and construct plots of the residuals versus fitted values and Q-Q plot of residuals to check your model assumptions. State which assumptions you are checking with each plot, and what are your conclusions.

   (b) Interpret the estimated regression coefficients. Suggest how you could make the regression coefficients more meaningful (if necessary). Make these changes, refit the model and interpret the new estimated coefficients.

4. **Clustering**

   Perform a cluster analysis to identify distinct groups (clusters) of vehicles (if any). Use the `kmeans` command in MATLAB to cluster the data into 2 clusters and briefly comment on the results.

5. **Out of sample predictions**

   Use your final model from question 2 (c) to calculate predicted values of fuel efficiency in the test sample. Hence calculate the mean square error for the test sample predictions, and compare this with the mean square error for training sample predictions. You are looking for signs of overfitting. Based on this, do you think your model should be simpler?